Bayesian Multivariate Process Modeling for Prediction of Forest Attributes

Andrew O. FINLEY, Sudipto BANERJEE,
Alan R. EK, and Ronald E. MCROBERTS

This article investigates multivariate spatial process models suitable for predicting multiple forest attributes using a multisource forest inventory approach. Such data settings involve several spatially dependent response variables arising in each location. Not only does each variable vary across space, they are likely to be correlated among themselves. Traditional approaches have attempted to model such data using simplifying assumptions, such as a common rate of decay in the spatial correlation or simplified cross-covariance structures among the response variables. Our current focus is to produce spatially explicit, tree species specific, prediction of forest biomass per hectare over a region of interest. Modeling such associations presents challenges in terms of validity of probability distributions as well as issues concerning identifiability and estimability of parameters. Our template encompasses several models with different correlation structures. These models represent different hypotheses whose tenability are assessed using formal model comparisons. We adopt a Bayesian hierarchical approach offering a sampling-based inferential framework using efficient Markov chain Monte Carlo methods for estimating model parameters.

Key Words: Bayesian inference; Coregionalization; Forest inventory; Markov chain Monte Carlo; Multivariate spatial process.

1. INTRODUCTION

Motivated by a need to produce spatially explicit predictions of multiple forest attributes using a multisource forest inventory approach, this manuscript develops a template for fitting a wide variety of multivariate Gaussian spatial process models. Specific interest lies in providing these predictions along with estimates of associated uncertainty for arbitrarily defined areas of interest (i.e., small-area prediction for any area in the domain of interest). This need is increasingly recognized by agencies conducting forest inventory and

Andrew O. Finley is an Assistant Professor in the Department of Forestry and Department of Geography, Michigan State University, East Lansing, MI, USA. Sudipto Banerjee is an Associate Professor in the Division of Biostatistics, School of Public Health, University of Minnesota, Minneapolis, MN. Alan R. Ek is Professor and Head at the Department of Forest Resources, University of Minnesota, Saint Paul, MN. Ronald E. McRoberts is a Mathematical Statistician at the North Central Research Station, USDA Forest Service, Saint Paul, MN. Corresponding author: Andrew O. Finley (E-mail: *finleya@msu.edu*).

Journal of Agricultural, Biological, and Environmental Statistics, Volume 13, Number 1, Pages 60–83
DOI: 10.1198/108571108X273160

in research based on small-area, high-intensity, inventories from experimental or research forests.

Small-area prediction of forest attributes and associated estimates of uncertainty are used to inform forest management decisions, further environmental research, and to serve as base data for a range of environmental monitoring initiatives. For instance, with the advent of the UN Framework Convention on Climate Change, and the subsequent Kyoto Protocol, came a need to map forest biomass and other variables related to measurements of current carbon stocks and flux. Consequently, several National Forest Inventory (NFI) programs have adapted their inventory and analysis to better support reporting and research on carbon budgets (e.g., the Finnish NFI and United States Forest Service Forest Inventory and Analysis program). Numerous studies, some of which are noted below, have found that multisource inventory methods coupling inventory plot data with remotely sensed imagery improve the prediction and mapping of forest biomass and other important economic and ecological forest variables.

Multisource forest inventory methods combine forest inventory field plot data with remotely sensed imagery, most commonly in the form of mid-resolution satellite imagery (e.g., from the Landsat sensors), to improve large- and small-scale estimates of forest attributes. Several approaches have been employed; among these, k-nearest neighbor (kNN) and traditional geostatistical methods are among the most popular. A variety of kNN methods have been proposed and successfully applied in several NFIs (Katila and Tomppo 2001; Trotter et al. 1997; McRoberts, Nelson, and Wendt 2002; Tomppo and Halme 2003). The kNN methods often provide useful point estimates (e.g., regional estimates of timber volume or biomass per hectare); however, Tomppo and Halme (2003) noted that, "the predictions and their standard errors computed from field data are only employed in validating all multisource predictions in areas ranging from several hundreds of thousand hectares to several million hectares. This is due to the fact that multisource error estimation for areas larger than a pixel (field plot) is complicated and the solution is yet to be found" (p. 99). The complication they referred to is the spatial autocorrelation among adjacent predictions (i.e., prediction for multi-pixel areas).

Geostatistical methods such as kriging and cokriging (see, e.g., Cressie 1993; Chilés and Delfiner 1999) have been used to account for spatial dependence in multisource inventory modeling (see Hudak et al. 2002; Lappi 2001). The spatial regression techniques used by Hudak et al. (2002) and the references therein, follow a traditional approach of modeling out spatial dependence structure using point estimates of semivariogram model parameters (e.g., nugget, sill, and range). These parameters are often highly variable as they are notoriously ill-defined by the data, especially the spatial range parameter. Because our focus is on predicting forest attributes for a small area, and most importantly a measure of uncertainty about those predictions, ignoring the potentially large variation in these parameters will result in falsely precise estimates. Furthermore, from a modeling standpoint, the traditional spatial regression models are narrow in scope, especially in multivariate settings.

From a statistical perspective, our primary goal is to incorporate rich correlation structures in models that are computationally feasible. For spatial data in general, and multivariate point-referenced data in particular, obtaining spatial correlation structures from descriptive methods such as empirical variograms is challenging at best. Indeed, in multivariate settings we envision a number of spatially varying dependent responses arising from each location. These variables are likely to be associated among themselves as well as across locations. Modeling such associations presents challenges in terms of validity of probability distributions as well as issues concerning identifiability and estimability. A more appealing way to identify suitable models is through a template encompassing several models with different correlation structures. These models represent different hypotheses whose tenability can be assessed using formal model comparisons. We adopt a Bayesian hierarchical approach (Gelman et al. 2003; Carlin and Louis 2000) for developing this template. Such an approach not only enables richer modeling but offers a richer inferential framework using samples from posterior distributions of model parameters. Similarly, predictions proceed from sampling the posterior predictive distribution that averages over uncertainty in parameter estimation.

The remainder of this article is organized as follows. Section 2 describes some features of our motivating dataset that couples forest inventory data from the USDA Forest Service Bartlett Experimental Forest with imagery from the Landsat sensor and other variables to map predicted forest biomass by tree species. Section 3 discusses spatial regression models arising in multivariate process contexts. Section 4 outlines the generalized template to implement these models and explains how we carry out inference and spatial predictions in a sampling-based framework. In Section 5 we return to the BEF data and explore several candidate models discussed in Section 4, provide parameter estimates for the “best” model, and offer maps of species specific predicted biomass per hectare with associated errors. Finally, Section 6 provides a summarizing discussion and indicates future work.

2. THE MOTIVATING DATASET

In an effort to better understand forest carbon dynamics in the Northeastern United States, total biomass by tree species is recorded on permanent forest inventory plots across the USDA Forest Service Bartlett Experimental Forest (BEF) in Bartlett, NH. The 1,053 hectare BEF covers a large elevation gradient from the village of Bartlett in the Saco River valley at 207 meters to about 914 meters above sea level (Figure 1). The BEF’s nearly continuous forest canopy is dominated by American beech (*Fagus grandifolia*), eastern hemlock (*Tsuga canadensis*), red maple (*Acer rubrum*), sugar maple (*Acer saccharum*), and yellow birch (*Betula alleghaniensis*). In 1931–1932, 500 permanent 0.1 hectare square inventory plots were located on a 200 × 100 meter grid. In the 2002 reinventory, 437 plots of the original 500 were georeferenced and remeasured.

Our central interest is to produce data layers of metric tons of above-ground biomass per hectare by tree species across the BEF. Because data layers such as these serve as input variables to subsequent models, it is crucial that each layer provides a spatially explicit measure of uncertainty.

As noted in the introduction, satellite imagery and other remotely sensed variables have proved useful regressors in multisource inventory of forest attributes such as biomass per

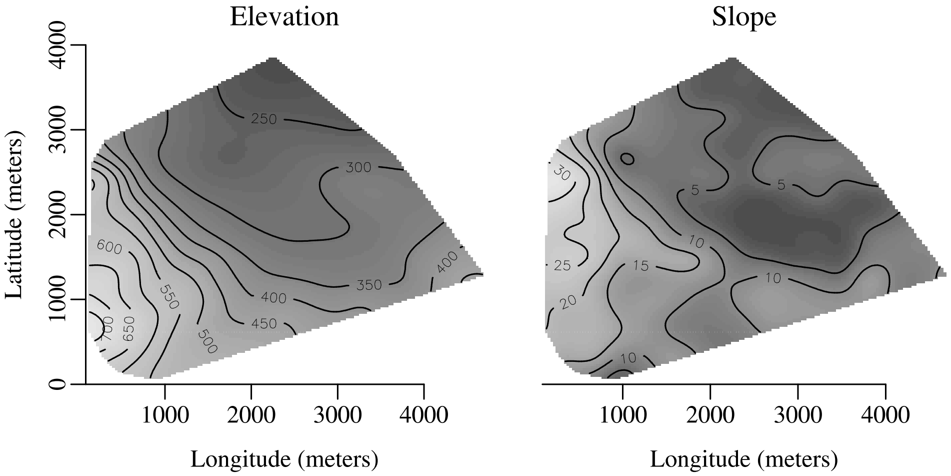

Figure 1. Elevation in meters above sea level and slope percent across the BEF.

hectare. Three dates of mid-resolution Landsat 7 ETM+ satellite imagery were acquired for the BEF. The imagery was obtained from the National Land Cover Database (*www.mrlc.gov/mrlc2k_nlcd.asp*). All images were geo-rectified to a common base layer, each with a root mean square error of less than 30 meters. In the rectification process, images were resampled to a 30 × 30 meter spatial resolution using the cubic convolution algorithm (see Homer et al. 2004). The three dates of imagery are April 14, 2003, August 9, 2002, and October 22, 2000, corresponding to early and peak vegetation green-up and senescence.

Each image was transformed to tasseled cap (TC) components of brightness (1), greenness (2), and wetness (3) using data-reduction techniques (Huang et al. 2002). The nine resulting spectral variables are referred to as AprTC1, AprTC2, AprTC3, AugTC1, AugTC2, AugTC3, and OctTC1, OctTC2, OctTC3. Thus, AprTC1 represents a spectral variable corresponding to the brightness level in April. In addition to these spectral variables, digital elevation model data was used to produce a 30 × 30 elevation (ELEV) and slope (SLOPE) layer for the BEF, Figure 1 (see *http://seamless.usgs.gov* for metadata). The centroids of the 437 georeferenced inventory plots were intersected with the elevation (ELEV), slope (SLOPE), and nine spectral variables.

As previously noted, five tree species comprise the plurality of forest cover on the BEF. Therefore, the plot variables of interest are estimated metric tons of above-ground biomass per hectare for American beech (BE), eastern hemlock (EH), red maple (RM), sugar maple (SM), and yellow birch (YB). Figure 2 provides an interpolated surface for each of the five response variables.

To demonstrate prediction, we randomly divided the plots into two sets of 218. The first set serves to fit the candidate models and the second *holdout* set is used for validation. We stress that this split-set approach is used here only for illustration and final inferences will be drawn from the full set of 437 plots.

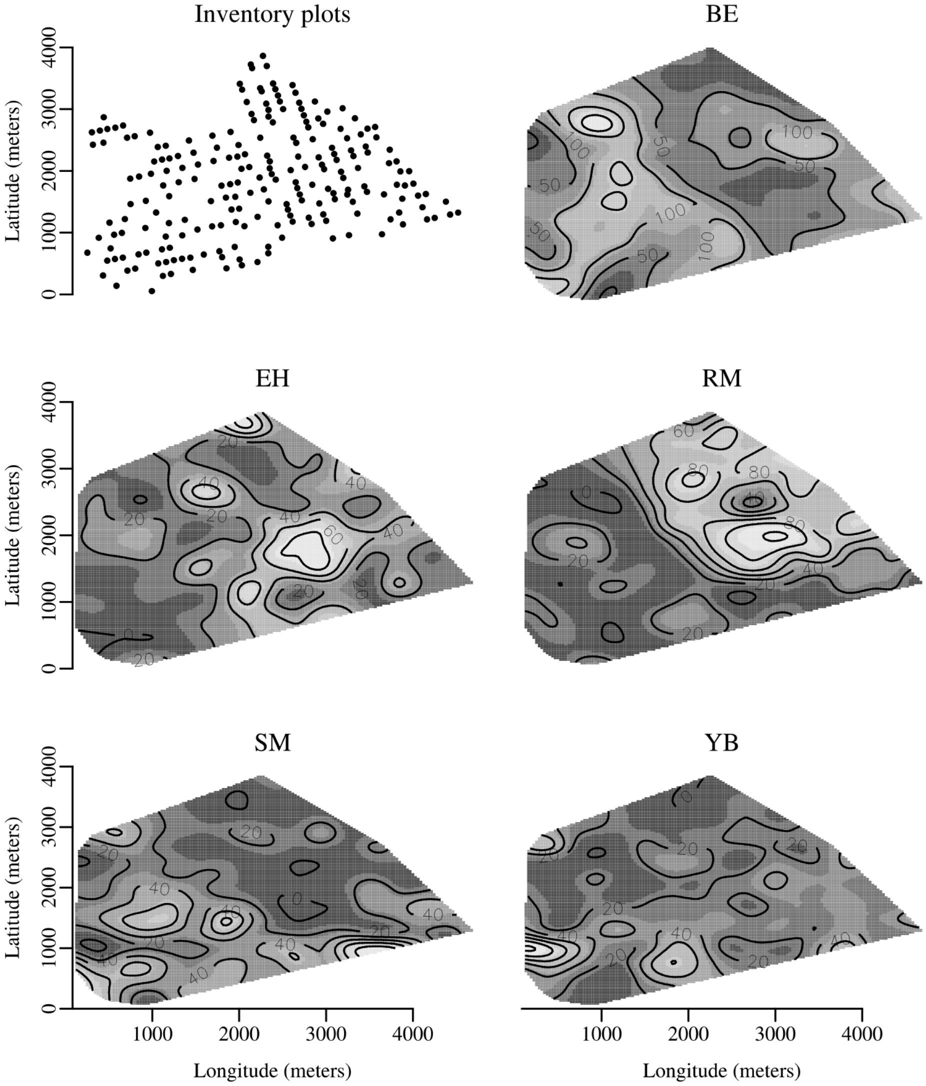

Figure 2. Interpolation of metric tons of biomass per hectare by species measured on inventory plots across the BEF. This set of 218 inventory plots was used for model parameter estimation.

3. MULTIVARIATE SPATIAL REGRESSION MODELS

The multivariate setting envisions a multivariate spatial regression model comprising an $m \times 1$ response vector $\mathbf{Y}(\mathbf{s}) = [Y_i(\mathbf{s})]_{i=1}^m$ along with an $m \times p$ ($p = \sum_{i=1}^m p_i$) matrix of regressors $\mathbf{X}^T(\mathbf{s}) = [\mathbf{x}_i^T(\mathbf{s})]_{i=1}^m$ connected through

$$\mathbf{Y}(\mathbf{s}) = \mathbf{X}^T(\mathbf{s})\boldsymbol{\beta} + \mathbf{W}(\mathbf{s}) + \boldsymbol{\epsilon}(\mathbf{s}), \tag{3.1}$$

where $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_p)^T$ is a $p \times 1$ vector of regression coefficients with $\boldsymbol{\beta}_i$ being the $p_i \times 1$ vector of regression coefficients corresponding to $\mathbf{x}_i^T(\mathbf{s})$, $\mathbf{W}(\mathbf{s}) \sim \text{MVGP}(\mathbf{0}, \mathbf{K}(\cdot, \cdot))$ is an $m \times 1$ zero-centered *multivariate Gaussian Process* capturing spatial variation, and $\boldsymbol{\epsilon}(\mathbf{s}) \sim \text{MVN}(\mathbf{0}, \Psi)$ models the measurement error effect for the response with the $m \times m$ dispersion matrix Ψ.

The critical ingredient for spatial modeling in (3.1) is the multivariate Gaussian process. We write an $m \times 1$ process as $\mathbf{W}(\mathbf{s}) \sim \text{MVGP}(\mathbf{0}, \mathbf{K}(\cdot, \cdot))$, where $\mathbf{W}(\mathbf{s}) = [W_i(\mathbf{s})]_{i=1}^m$ is an $m \times 1$ vector process with an $m \times m$ *cross-covariance* matrix function $\mathbf{K}(\mathbf{s}, \mathbf{s}') = [\text{cov}(W_i(\mathbf{s}), W_j(\mathbf{s}'))]_{i,j=1}^m$ whose (i, j)th element is the covariance function between $W_i(\mathbf{s})$ and $W_j(\mathbf{s}')$. For any integer n and any collection of sites $\mathbf{s}_1, \ldots, \mathbf{s}_n$, we write the multivariate realizations as an $mn \times 1$ vector $\mathbf{W} = (\mathbf{W}^T(\mathbf{s}_1), \ldots, \mathbf{W}^T(\mathbf{s}_n))^T$ which is distributed as an $mn \times 1$ multivariate normal distribution: $\mathbf{W} \sim \text{MVN}(\mathbf{0}, \Sigma_{\mathbf{W}})$, where $\Sigma_{\mathbf{W}} = [\mathbf{K}(\mathbf{s}_i, \mathbf{s}_j)]_{i,j=1}^n$ is an $mn \times mn$ matrix that can be partitioned as an $n \times n$ block matrix comprising $m \times m$ blocks with the (i, j)th block being the cross-covariance matrix $\mathbf{K}(\mathbf{s}_i, \mathbf{s}_j)$. Letting $\mathbf{Y} = [\mathbf{Y}(\mathbf{s}_i)]_{i=1}^n$ be the $mn \times 1$ observed response vector, its dispersion matrix becomes $\Sigma_{\mathbf{Y}} = \Sigma_{\mathbf{W}} + I_n \otimes \Psi$, where I_n is the $n \times n$ identity matrix and $\otimes$ denotes the Kronecker product (e.g., Harville 1997).

Clearly, care is needed in choosing $\mathbf{K}(\cdot, \cdot)$ so that $\Sigma_{\mathbf{W}}$ is symmetric and positive definite. Characterizing valid cross-covariance matrix functions that ensure positive-definiteness of $\Sigma_{\mathbf{W}}$ is indeed more demanding than the choice of real-valued covariance functions in univariate spatial modeling that are characterized by Bochner's Theorem (see, e.g., Cressie 1993). In the multivariate setting, we require that for an arbitrary number and choice of locations the resulting $\Sigma_{\mathbf{W}}$ be symmetric and positive definite. In fact, note that the cross-covariance matrix function need not be symmetric or positive definite but must satisfy $\mathbf{K}(\mathbf{s}', \mathbf{s}) = \mathbf{K}^T(\mathbf{s}, \mathbf{s}')$ so that $\Sigma_{\mathbf{W}}$ is symmetric. In the limiting sense, as $\mathbf{s} \to \mathbf{s}'$, $\mathbf{K}(\mathbf{s}, \mathbf{s})$ does become symmetric and positive definite as it models the covariances between the different components of $\mathbf{W}(\mathbf{s})$ *within* site $\mathbf{s}$. A theorem by Cramér (see, e.g., Chilés and Delfiner 1999) provides a characterization of cross-covariance functions, akin to Bochner's theorem for covariance functions, but using Cramér's result in practical modeling is less trivial. Majumdar and Gelfand (2006) offered a review of other approaches such as convolution of covariance functions that lead to valid cross-covariances, but they too recognized the computational and modeling difficulties involved.

Since our primary objective is to develop a computationally feasible template that accommodates sufficiently rich multivariate spatial models, we adopt a constructive approach that has recently gained popularity through *coregionalization* models (Wackernagel 2003). To motivate this approach, consider how some of the simplest cross-covariance functions arise. For instance, suppose that $\tilde{\mathbf{W}}(\mathbf{s}) = [\tilde{W}_i(\mathbf{s})]_{i=1}^m$ is an $m \times 1$ process with independent zero-centered spatial processes with unit variance; that is, each $\tilde{W}_k(\mathbf{s}) \sim GP(0, \rho_k(\cdot, \cdot))$ with $\text{var}(\tilde{W}_k(\mathbf{s})) = 1$ and $\text{cov}(\tilde{W}_k(\mathbf{s}), \tilde{W}_k(\mathbf{s}')) = \rho_k(\mathbf{s}, \mathbf{s}'; \boldsymbol{\theta}_k)$, where $\rho_k(\cdot; \boldsymbol{\theta}_k)$ is a correlation function associated with $\tilde{W}_k(\mathbf{s})$ and $\boldsymbol{\theta}_k$ are parameters therein. Also note that $\text{cov}(\tilde{W}_k(\mathbf{s}), \tilde{W}_k'(\mathbf{s}')) = 0$ when $k \neq k'$ (irrespective of how close $\mathbf{s}$ and $\mathbf{s}'$ are), which implies that the cross-covariance matrix function $\tilde{\mathbf{K}}(\mathbf{s}, \mathbf{s}'; \boldsymbol{\theta})$ with $\boldsymbol{\theta} = \{\boldsymbol{\theta}_k\}_{k=1}^m$ is simply a diagonal matrix with (k, k)th element being $\rho_k(\mathbf{s}, \mathbf{s}'; \boldsymbol{\theta}_k)$. It easily follows that $\tilde{\mathbf{K}}(\mathbf{s}, \mathbf{s}'; \boldsymbol{\theta})$

is a valid cross-covariance matrix since each of its diagonal elements is a valid real-valued positive definite function.

The Matérn correlation function allows control of spatial association and smoothness (see, e.g., Stein 1999) and is given by

$$\rho(\mathbf{s},\mathbf{s}';\phi,\nu)=\frac{1}{2^{\nu-1}\Gamma(\nu)}(\|\mathbf{s}-\mathbf{s}'\|\phi)^{\nu}\mathscr{K}_{\nu}(\|\mathbf{s}-\mathbf{s}'\|;\phi);\quad \phi>0,\ \nu>0, \tag{3.2}$$

where ϕ controls the decay in spatial correlation and ν is a smoothness parameter with higher values yielding smoother process realizations. Also, Γ is the usual Gamma function while $\mathscr{K}_{\nu}$ is a modified Bessel function of the third kind with order ν and $\|\mathbf{s}-\mathbf{s}'\|$ is the Euclidean distance between the sites $\mathbf{s}$ and $\mathbf{s}'$. Covariance functions that depend only on the distance metric are often referred to as *isotropic*. Several other choices for valid correlation functions were discussed by Banerjee et al. (2004). For $\tilde{W}_k(\mathbf{s})$ we choose isotropic Matérn functions $\rho_k(\mathbf{s},\mathbf{s}';\boldsymbol{\theta}_k)$ with $\boldsymbol{\theta}_k=(\phi_k,\nu_k)$ for $k=1,\ldots,m$. Note that $\rho_k(\mathbf{s},\mathbf{s}')$ is the correlation function for the kth component of $\tilde{\mathbf{W}}(\mathbf{s})$ and does not correspond to the kth component of the observed vector process $\mathbf{Y}(\mathbf{s})$. Consequently, the parameters ν_k and ϕ_k do not correspond directly to $\mathbf{Y}(\mathbf{s})$, but to the unobserved process $\tilde{\mathbf{W}}(\mathbf{s})$ that drives the spatial variation in $\mathbf{Y}(\mathbf{s})$.

For building richer covariance structures, we assume the process $\mathbf{W}(\mathbf{s})=\mathbf{A}(\mathbf{s})\tilde{\mathbf{W}}(\mathbf{s})$ to be a linear transformation of $\tilde{\mathbf{W}}(\mathbf{s})$, where $\mathbf{A}(\mathbf{s})$ is a space-varying matrix transform that is nonsingular for all $\mathbf{s}$. Then, the cross-covariance matrix functions are related as $\mathbf{K}(\mathbf{s},\mathbf{s}';\boldsymbol{\theta})=\mathbf{A}(\mathbf{s})\tilde{\mathbf{K}}(\mathbf{s},\mathbf{s}')\mathbf{A}^T(\mathbf{s}')$. It is worth noting that $\tilde{\mathbf{K}}(\mathbf{s},\mathbf{s};\boldsymbol{\theta})=I_m$ (the $m\times m$ identity matrix), so that $\mathbf{K}(\mathbf{s},\mathbf{s};\boldsymbol{\theta})=\mathbf{A}(\mathbf{s})\mathbf{A}^T(\mathbf{s})$. Therefore $\mathbf{A}(\mathbf{s})=\mathbf{K}^{1/2}(\mathbf{s},\mathbf{s};\boldsymbol{\theta})$ is identified as a square root (e.g. Cholesky) of $\mathbf{K}(\mathbf{s},\mathbf{s};\boldsymbol{\theta})$ and can be taken to be lower-triangular without loss of generalization. Indeed, the one-to-one correspondence between the elements of the square-root matrix and the original matrix is well known (see, e.g., Harville 1997, p. 229). The validity of $\mathbf{K}(\mathbf{s},\mathbf{s}';\boldsymbol{\theta})$ follows immediately from that of $\tilde{\mathbf{K}}(\mathbf{s},\mathbf{s}';\boldsymbol{\theta})$ and the dispersion matrix of realizations of $\mathbf{W}(\mathbf{s})$, $\Sigma_{\mathbf{W}}=[\mathbf{K}(\mathbf{s}_i,\mathbf{s}_j;\boldsymbol{\theta})]_{i,j=1}^n$ can be written as

$$\begin{aligned}[\mathbf{A}(\mathbf{s}_i)\tilde{\mathbf{K}}(\mathbf{s}_i,\mathbf{s}_j;\boldsymbol{\theta})\mathbf{A}^T(\mathbf{s}_j)]_{i,j=1}^n &= [\oplus_{i=1}^k\mathbf{A}(\mathbf{s}_i)][\oplus_{k=1}^m\rho_k(\mathbf{s}_i,\mathbf{s}_j;\boldsymbol{\theta}_k)]_{i,j=1}^n[\oplus_{i=1}^k\mathbf{A}^T(\mathbf{s}_i)]\\ &= \mathscr{A}\Sigma_{\tilde{\mathbf{W}}}\mathscr{A}^T, \end{aligned}\tag{3.3}$$

where $\oplus$ is the "diagonal" or direct-sum matrix operator (e.g., Harville 1997) so, for each (i,j), $\oplus_{k=1}^m\rho_k(\mathbf{s}_i,\mathbf{s}_j;\boldsymbol{\theta}_k)$ is an $m\times m$ diagonal matrix with $\rho_k(\mathbf{s}_i,\mathbf{s}_j;\boldsymbol{\theta})$ as its diagonals while $\mathscr{A}$ is a block-diagonal matrix with the ith diagonal block being $\mathbf{A}(\mathbf{s}_i)$. This model is essentially a spatially adaptive version of the linear model of coregionalization (LMC) in the geostatistics literature (Wakernagel 2003; Gelfand et al. 2004; Zhang 2006). This is a highly structured model that models the cross-covariance function as $\mathbf{K}(\mathbf{s},\mathbf{s}')=\sum_{k=1}^m\mathbf{a}_k(\mathbf{s})\mathbf{a}_k^T(\mathbf{s}')\rho_k(\mathbf{s},\mathbf{s}')$, where $\mathbf{a}_k(\mathbf{s})$ is the kth column vector of $A(\mathbf{s})$.

Note that the space-varying linear transformation $\mathbf{A}(\mathbf{s})$ induces a nonstationary process that might often be realistic and yield better estimates. However, treating site as completely unknown may create problems as one would need to assign space-varying priors on them; for example, we could treat each element of $\mathbf{A}(\mathbf{s})$ as a spatial process or construct an inverted-Wishart process as in Gelfand et al. (2004). Realistically, though, we would rarely find data that contains enough information to estimate such processes. In certain experimental settings with nested sites, such as in forest inventories or agricultural experiments, it is possible to model $\mathbf{A}(\mathbf{s})$ assuming an embedded spatial process within site $\mathbf{s}$, leading to multiresolution spatial models. See, for example, Banerjee and Johnson (2006) and Banerjee and Finley (2007) for such applications.

Stationary cross-covariance functions, on the other hand, necessarily imply the linear transformation to be independent of space. Here, since the cross-covariance is a function of the separation between sites, we have $\mathbf{K}(\mathbf{s}, \mathbf{s}; \boldsymbol{\theta}) = \mathbf{K}(\mathbf{0}; \boldsymbol{\theta})$ so that $\mathbf{A}(\mathbf{s}) = \mathbf{A} = \mathbf{K}^{1/2}(\mathbf{0}; \boldsymbol{\theta})$. In such cases, $\mathscr{A} = I \otimes A$ and (3.3) reduces to

$$\Sigma_{\mathbf{W}} = (I_n \otimes A)\Sigma_{\tilde{\mathbf{W}}}(I_n \otimes A^T). \tag{3.4}$$

As a further simplification, suppose we choose $\tilde{\mathbf{K}}(\mathbf{s}, \mathbf{s}'; \boldsymbol{\theta}) = \rho(\mathbf{s}-\mathbf{s}'; \boldsymbol{\theta})I_m$, that is, a single correlation function for each component of $\tilde{\mathbf{W}}(\mathbf{s})$. This yields $\Sigma_{\tilde{\mathbf{W}}} = R(\boldsymbol{\theta}) \otimes I_m$, where $R(\boldsymbol{\theta}) = [\rho(\mathbf{s}_i, \mathbf{s}_j; \boldsymbol{\theta})]_{i,j=1}^n$ and results in a *separable* or *intrinsic* specification (see, e.g., Wackernagel 2003):

$$\Sigma_{\mathbf{W}} = (I_n \otimes A)(R \otimes I_m)(I_n \otimes A^T) = R(\boldsymbol{\theta}) \otimes \mathbf{K}(\mathbf{0}; \boldsymbol{\theta}). \tag{3.5}$$

Here, the dispersion structure separates into a spatial component $R(\boldsymbol{\theta})$ and a within-site dispersion matrix $\mathbf{K}(\mathbf{0}; \boldsymbol{\theta})$. While such models have nicer interpretability, they are often too simplistic and provide poorer fits to the data.

4. BAYESIAN IMPLEMENTATION USING A GENERALIZED TEMPLATE

4.1 ESTIMATION OF MODEL PARAMETERS

We adopt a Bayesian approach specifying prior distributions on the parameters to build hierarchical models that are estimated using a Gibbs sampler, with Metropolis updates when required, for fitting our models (see, e.g., Gelman et al. 2003; chap. 11). Although such algorithms are usually problem-specific, often requiring intensive coding, casting the problem in a general template allows several models to be fit without rewriting vast amounts of code. We cast the data model into the following generic template:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathscr{A}\tilde{\mathbf{W}} + \boldsymbol{\epsilon};\ \boldsymbol{\epsilon} \sim N(\mathbf{0}, I_m \otimes \Psi), \tag{4.1}$$

where $\mathbf{Y}$ is the $mn \times 1$ response vector, and $\mathbf{X}$ is the $mn \times p$ matrix of regressors, $\boldsymbol{\beta}$ is the corresponding vector of regression coefficients. The specifications for the $mn \times mn$ matrix $\mathscr{A}$ and $\tilde{\mathbf{W}}$ give rise to different multivariate spatial regression models. Markov chain Monte Carlo (MCMC) model fitting proceeds with a Gibbs sampler with Metropolis steps (see, e.g., Gelman et al. 2003) on the marginalized scale, after integrating out $\tilde{\mathbf{W}}$, to reduce the parameter space. The marginalized likelihood becomes $\text{MVN}(\mathbf{X}\boldsymbol{\beta}, \mathscr{A}\Sigma_{\tilde{\mathbf{W}}}\mathscr{A}^T + I_n \otimes \Psi)$.

Bayesian hierarchical models are completed by assigning prior distributions on the parameters. Customarily, we set $\boldsymbol{\beta} \sim \mathrm{MVN}(\boldsymbol{\mu}_{\boldsymbol{\beta}}, \Sigma_{\boldsymbol{\beta}})$ to a p-dimensional multivariate normal distribution, while the measurement error dispersion Ψ could be assigned an inverse-Wishart prior although one usually assumes independence of measurement error for the different response measurements in each site and sets $\Psi = \mathrm{diag}(\tau_i^2)_{i=1}^m$ as a diagonal matrix with each $\tau_i^2 \sim \mathrm{IG}(a_i, b_i)$. Also recall that $\mathscr{A}$ itself is unknown and needs to be stochastically specified. As mentioned in Section 3, the specific form of $\mathscr{A}$ will depend upon the exact form of $\mathbf{A}$. For the stationary setting, we have $\mathscr{A} = I_n \otimes \mathbf{A}$ and we assign an inverse-Wishart prior to $\mathbf{A}\mathbf{A}^T$. Finally recall that $\Sigma_{\tilde{\mathbf{W}}} = [\tilde{\mathbf{K}}(\mathbf{s}_i - \mathbf{s}_j; \boldsymbol{\theta})]_{i,j=1}^n$ and one needs to assign priors on $\boldsymbol{\theta} = \{\phi_k, \nu_k\}_{k=1}^m$. This will again depend upon the specific choice of the correlation functions. In general the spatial decay parameters are weakly identifiable and prior specifications become an even more delicate issue. Reasonably informative priors are needed for satisfactory MCMC behavior and typically we set prior distributions for the decay parameters relative to the size of their domains, for instance by setting the prior means to values that imply the spatial ranges to approximately a certain fraction of the maximum distance. For the Matérn correlation function, the smoothness parameter ν is often estimated using a $U(0, 2)$ prior distribution. This choice is motivated by earlier findings (e.g., Stein 1999) that it is almost impossible for the data to distinguish between these smoothness parameters for values greater than 2.

Generically denoting by $\Omega = (\boldsymbol{\beta}, \mathscr{A}, \boldsymbol{\theta}, \Psi)$ the set of parameters that are to be updated in the marginalized model from (4.1), we need to sample from the posterior distribution

$$P(\Omega \,|\, \mathrm{Data}) \propto P(\boldsymbol{\beta})P(\mathscr{A})P(\boldsymbol{\theta})P(\Psi)P(\mathbf{Y} \,|\, \boldsymbol{\beta}, \mathscr{A}, \boldsymbol{\theta}, \Psi). \tag{4.2}$$

An efficient MCMC algorithm is obtained by updating $\boldsymbol{\beta}$ from its full conditional $\mathrm{MVN}(\boldsymbol{\mu}_{\boldsymbol{\beta}|\cdot}, \Sigma_{\boldsymbol{\beta}|\cdot})$, where

$$\Sigma_{\boldsymbol{\beta}|\cdot} = [\Sigma_{\boldsymbol{\beta}}^{-1} + \mathbf{X}^T(\mathscr{A}\Sigma_{\tilde{\mathbf{W}}}\mathscr{A}^T + I_n \otimes \Psi)^{-1}\mathbf{X}]^{-1};$$
$$\boldsymbol{\mu}_{\boldsymbol{\beta}|\cdot} = \Sigma_{\boldsymbol{\beta}|\cdot}\mathbf{X}^T(\mathscr{A}\Sigma_{\tilde{\mathbf{W}}}\mathscr{A}^T + I_n \otimes \Psi)^{-1}\mathbf{Y}.$$

All the remaining parameters have to be updated using Metropolis–Hastings steps. Depending upon the application, this may be implemented using block-updates (e.g., all the parameters in Ψ in one block and those in $\mathscr{A}$ in another). On convergence, the MCMC output generates L samples, say $\{\Omega^{(l)}\}_{l=1}^L$, from the posterior distribution in (4.2).

4.2 Posterior Predictive Inference

In updating Ω using the marginal model as outlined above, we do not directly sample the spatial coefficients $\tilde{\mathbf{W}}$ and hence cannot directly obtain $\mathbf{W} = \mathscr{A}\tilde{\mathbf{W}}$. This shrinks the parameter space resulting in a more efficient MCMC algorithm. A primary advantage of the first-stage Gaussian models (as in (4.1)) is that the posterior distribution of $\tilde{\mathbf{W}}$ can be recovered in a posterior predictive fashion by sampling from

$$P(\tilde{\mathbf{W}}|\ \mathrm{Data}) \propto \int P(\tilde{\mathbf{W}}|\Omega,\ \mathrm{Data})P(\Omega|\ \mathrm{Data})d\Omega. \tag{4.3}$$

Once the posterior samples from $P(\Omega|\text{ Data})$, $\{\Omega^{(l)}\}_{l=1}^{L}$, have been obtained, posterior samples from $P(\tilde{\mathbf{W}}|\text{ Data})$ are drawn by sampling $\tilde{\mathbf{W}}^{(l)}$ for each $\boldsymbol{\Omega}^{(l)}$ from $P(\tilde{\mathbf{W}}\,|\,\Omega^{(l)},\text{ Data})$. This composition sampling is routine because $P(\tilde{\mathbf{W}}\,|\,\Omega,\text{ Data})$ in (4.3) is Gaussian; in fact, from (4.1) we have this distribution as

$$\text{MVN}\Big[(\Sigma_{\tilde{\mathbf{W}}}^{-1} + \mathscr{A}^T(I_n \otimes \Psi^{-1})\mathscr{A})^{-1}\mathscr{A}^T(I_n \otimes \Psi^{-1})(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}),$$
$$(\Sigma_{\tilde{\mathbf{W}}}^{-1} + \mathscr{A}^T(I_n \otimes \Psi^{-1})\mathscr{A})^{-1}\Big].$$

The posterior estimates of these realizations can subsequently be mapped with contours to produce image and contour plots of the spatial processes.

Let $\{\mathbf{s}_{0i}\}_{i=1}^{n^*}$ be a collection of n^* locations where we seek to predict the response. It might also be of interest to compute the posterior predictive distribution $P(\tilde{\mathbf{W}}^*\,|\,\text{Data})$ where $\tilde{\mathbf{W}}^* = [\tilde{\mathbf{W}}(\mathbf{s}_{0k})]_{k=1}^{n^*}$. Note that

$$P(\tilde{\mathbf{W}}^*|\text{ Data}) \propto \int P(\tilde{\mathbf{W}}^*|\tilde{\mathbf{W}},\Omega,\text{ Data})P(\tilde{\mathbf{W}}|\Omega,\text{ Data})P(\Omega|\text{ Data})d\Omega d\tilde{\mathbf{W}}. \qquad (4.4)$$

This can be computed by composition sampling by first obtaining the posterior samples $\{\Omega^{(l)}\}_{l=1}^{L} \sim P(\Omega|\text{ Data})$, then drawing $\tilde{\mathbf{W}}^{(l)} \sim P(\tilde{\mathbf{W}}|\Omega^{(l)},\text{ Data})$ for each l as described in (4.3) and finally drawing $\tilde{\mathbf{W}}^{*(l)} \sim P(\tilde{\mathbf{W}}^*|\tilde{\mathbf{W}}^{(l)}\Omega^{(l)},\text{ Data})$. This last distribution is derived as a conditional distribution from a multivariate normal distribution as follows:

$$\begin{pmatrix}\tilde{\mathbf{W}}\\ \tilde{\mathbf{W}}^*\end{pmatrix} \sim \text{MVN}\left(\begin{pmatrix}\mathbf{0}\\ \mathbf{0}\end{pmatrix}, \begin{pmatrix}\Sigma_{\tilde{\mathbf{W}}} & \Sigma_{\tilde{\mathbf{W}},\tilde{\mathbf{W}}^*}\\ \Sigma_{\tilde{\mathbf{W}}^*,\tilde{\mathbf{W}}} & \Sigma_{\tilde{\mathbf{W}}^*}\end{pmatrix}\right),$$
$$\begin{aligned}\text{where } \Sigma_{\tilde{\mathbf{W}}} &= [\oplus_{k=1}^{m}\rho_k(\mathbf{s}_i,\mathbf{s}_j;\boldsymbol{\theta}_k)]_{i,j=1}^{n}\\ \Sigma_{\tilde{\mathbf{W}}^*} &= [\oplus_{k=1}^{m}\rho_k(\mathbf{s}_{0i},\mathbf{s}_{0j};\boldsymbol{\theta}_k)]_{i,j=1}^{n^*},\\ \text{and } \Sigma_{\mathbf{W},\mathbf{W}^*}^T &= \Sigma_{\tilde{\mathbf{W}}^*,\tilde{\mathbf{W}}} = [\oplus_{k=1}^{m}\rho_k(\mathbf{s}_{0i},\mathbf{s}_j;\boldsymbol{\theta}_k)]_{i=1,j=1}^{n^*,n}.\end{aligned}$$

Therefore, the distribution $P(\tilde{\mathbf{W}}^*|\tilde{\mathbf{W}},\Omega,\text{ Data})$ is $\text{MVN}(\boldsymbol{\mu}_{\tilde{\mathbf{W}}^*|\tilde{\mathbf{W}}},\Sigma_{\tilde{\mathbf{W}}^*|\tilde{\mathbf{W}}})$, where

$$\begin{aligned}\boldsymbol{\mu}_{\tilde{\mathbf{W}}^*|\tilde{\mathbf{W}}} &= \Sigma_{\tilde{\mathbf{W}},\tilde{\mathbf{W}}^*}^T\Sigma_{\tilde{\mathbf{W}}}^{-1}\tilde{\mathbf{W}},\\ \Sigma_{\tilde{\mathbf{W}}^*|\tilde{\mathbf{W}}} &= \Sigma_{\tilde{\mathbf{W}}^*} - \Sigma_{\tilde{\mathbf{W}},\tilde{\mathbf{W}}^*}^T\Sigma_{\tilde{\mathbf{W}}}^{-1}\Sigma_{\tilde{\mathbf{W}},\tilde{\mathbf{W}}^*}.\end{aligned}$$

Once $\{\tilde{\mathbf{W}}^{*(l)}\}_{l=1}^{L}$ have been obtained, we can easily predict the responses, say $\mathbf{Y}^* = [\mathbf{Y}(\mathbf{s}_{0i})]_{i=1}^{n^*}$ at those sites as long as the $mn^* \times p$ matrix of regressors for those locations, say $\mathbf{X}^*$, is available. This can be done by simply sampling the conditional expectations $E[\mathbf{Y}^*|\text{ Data}]^{(l)} = \mathbf{X}^*\boldsymbol{\beta}^{(l)} + \mathscr{A}^{(l)}\tilde{\mathbf{W}}^{*l}$ for $l = 1,\ldots,L$. Equivalently, predictions can be executed by drawing posterior samples from the marginal distribution below, without resorting to direct updates of the $\tilde{\mathbf{W}}$ as follows:

$$P(\mathbf{Y}^*|\text{ Data}) \propto \int P(\mathbf{Y}^*|\Omega,\text{ Data})P(\Omega|\text{ Data})d\Omega. \qquad (4.5)$$

In the stationary setting with the marginalized model, observe that

$$\begin{pmatrix} \mathbf{Y} \\ \mathbf{Y}^* \end{pmatrix} \sim \text{MVN}\left(\begin{pmatrix} \mathbf{X}\boldsymbol{\beta} \\ \mathbf{X}^*\boldsymbol{\beta} \end{pmatrix}, \begin{pmatrix} \Sigma_{\mathbf{Y},\mathbf{Y}} & \Sigma_{\mathbf{Y},\mathbf{Y}^*} \\ \Sigma_{\mathbf{Y},\mathbf{Y}^*}^T & \Sigma_{\mathbf{Y}^*,\mathbf{Y}^*} \end{pmatrix}\right),$$

$$\begin{aligned} \text{where } \Sigma_{\mathbf{Y},\mathbf{Y}} &= \mathscr{A}\Sigma_{\tilde{\mathbf{W}}}\mathscr{A}^T + I_n \otimes \Psi \text{ with } \mathscr{A} = I_n \otimes \mathbf{A}, \\ \Sigma_{\mathbf{Y}^*,\mathbf{Y}^*} &= \mathscr{A}^*\Sigma_{\tilde{\mathbf{W}}^*}\mathscr{A}^{*T} \text{ with } \mathscr{A}^* = (I_{n^*} \otimes \mathbf{A}), \\ \text{and } \Sigma_{\mathbf{Y},\mathbf{Y}^*}^T &= \mathscr{A}^*\Sigma_{\tilde{\mathbf{W}}^*,\tilde{\mathbf{W}}}\mathscr{A}^T. \end{aligned}$$

Therefore, the distribution $P(\mathbf{Y}^*|\Omega, \text{ Data})$ is $\text{MVN}(\boldsymbol{\mu}_{\mathbf{Y}^*|\mathbf{Y}}, \Sigma_{\mathbf{Y}^*|\mathbf{Y}})$, where

$$\begin{aligned} \boldsymbol{\mu}_{\mathbf{Y}^*|\mathbf{Y}} &= \mathbf{X}^*\boldsymbol{\beta} + \Sigma_{\mathbf{Y},\mathbf{Y}^*}^T\Sigma_{\mathbf{Y},\mathbf{Y}}^{-1}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}), \\ \Sigma_{\mathbf{Y}^*|\mathbf{Y}} &= \Sigma_{\mathbf{Y}^*,\mathbf{Y}^*} - \Sigma_{\mathbf{Y},\mathbf{Y}^*}^T\Sigma_{\mathbf{Y},\mathbf{Y}}^{-1}\Sigma_{\mathbf{Y},\mathbf{Y}^*}. \end{aligned}$$

Simulating from $P(\mathbf{Y}^*|\Omega, \text{ Data})$ is routine for any given $\boldsymbol{\Omega}$. Hence, the predictive distribution is again obtained using composition sampling: for each $\boldsymbol{\Omega}^{(l)} \sim P(\Omega \mid \text{Data})$, we draw $\mathbf{Y}^{*,(l)} \sim P(\mathbf{Y}^*|\Omega^{(l)}, \text{ Data})$ to obtain posterior predictive samples $\{\mathbf{Y}^{*,l}\}_{l=1}^{L}$.

5. BEF DATA ANALYSIS

5.1 CANDIDATE MODELS

The generalized template introduced in Section 4 suggests several potential models for the BEF biomass data. Here we consider four stationary process models of increasing complexity. Our focus is on the alternative specifications of $\mathscr{A}$ and $\tilde{\mathbf{W}}$ within (4.1). For the candidate models, we assume independent response specific measurement error, $\Psi = \text{diag}(\tau_i^2)_{i=1}^m$, a common set of regressors (i.e., common $\mathbf{X}$ matrix), and an isotropic spatial process that can be modeled with the Matérn correlation function given in (3.2).

A simple linear regression model (no random effects) is obtained from setting

$$\text{Model 1: } \mathscr{A}\tilde{\mathbf{W}} = 0$$

in (4.1) and would suffice with negligible extraneous variation beyond what is explained by the regressors. However, we expect similar responses in proximate locations, possibly resulting from similar topographic and environmental conditions. This autocorrelation is seen in each response variable's surface (Figure 2). If the regressors do not account for correlation as a function of distance between locations, then this model violates the implicit assumption of conditionally independent observations.

The first two spatial models impose separable correlation structures as in (3.5). For each model, $\Sigma_{\tilde{\mathbf{W}}} = [\tilde{\mathbf{K}}(\mathbf{s}_i - \mathbf{s}_j; \boldsymbol{\theta})]_{i,j=1}^n$, $\boldsymbol{\theta} = \{\phi, \nu\}_{k=1}^m$ implies the response variables share a common spatial decay, ϕ, and smoothness parameter, ν. The first of these models assumes independently varying spatial processes,

$$\text{Model 2: } \mathscr{A} = I_n \otimes \text{diag}(\sigma_i)_{i=1}^m,$$

whereas the second attempts to capture the spatial covariances among the response variables within a location,

$$\text{Model 3: } \mathscr{A} = I_n \otimes A.$$

The next two models we investigate are nonseparable extensions of Models 2 and 3. Specifically, these models allow for process specific spatial decay and smoothness parameters, $\boldsymbol{\theta} = \{\phi_k, \nu_k\}_{k=1}^m$, for each of the components of $\tilde{\mathbf{W}}(\mathbf{s})$. The first of these, like Model 2, ignores the within location dependence between the response variables

$$\text{Model 4: } \mathscr{A} = I_n \otimes \text{diag}(\sigma_i)_{i=1}^m \text{ and } \Sigma_{\tilde{\mathbf{W}}} = [\tilde{\mathbf{K}}(\mathbf{s}_i - \mathbf{s}_j; \boldsymbol{\theta})]_{i,j=1}^n,$$

while the second one, like Model 5, allows for such dependence,

$$\text{Model 5: } \mathscr{A} = I_n \otimes A \text{ and } \Sigma_{\tilde{\mathbf{W}}} = [\tilde{\mathbf{K}}(\mathbf{s}_i - \mathbf{s}_j; \boldsymbol{\theta})]_{i,j=1}^n.$$

5.2 MODEL SELECTION

Since we consider several alternative models with varying degrees of spatial richness, we use the Deviance Information Criterion (DIC) (Spiegelhalter et al. 2002) as a measure of model choice. The DIC has nice properties for Gaussian likelihoods (as ours) and is particularly convenient to compute from posterior samples. This criterion is the sum of the Bayesian deviance (a measure of model fit) and the (effective) number of parameters (a penalty for model complexity). It rewards better fitting models through the first term and penalizes more complex models through the second term, with lower values indicating favorable models for the data. The deviance, up to an additive quantity not depending upon Ω, is simply the negative of twice the log-likelihood, $D(\Omega) = -2 \log L(\text{Data}|\ \Omega)$, where $L(\text{Data}|\ \Omega)$ is the first stage Gaussian likelihood from (4.1) for the respective models. The Bayesian deviance is the posterior mean, $\overline{D(\Omega)} = E_{\Omega|\mathbf{Y}}[D(\Omega)]$, while the effective number of parameters is given by $p_D = \overline{D(\Omega)} - D(\bar{\Omega})$, where $\bar{\Omega}$ is the posterior mean of the model parameters Ω. The DIC is then given by $\overline{D(\Omega)} + p_D$ and is easily computed from the posterior samples.

5.3 MODEL PARAMETER ESTIMATION

As noted in Section 4.1, the Bayesian hierarchical models are completed by assigning prior distribution to parameters $\Omega = (\boldsymbol{\beta}, \mathscr{A}, \boldsymbol{\theta}, \Psi)$. For each model, a flat prior was assigned to the regressor parameters in $\boldsymbol{\beta}$. The prior distributions for the remaining parameters are consistent with definitions found in Appendix A of Gelman et al. (2003). As noted earlier, we assume $\Psi = \text{Diag}(\tau_i^2)_{i=1}^5$ with each τ_i^2 receiving an inverse-Gamma prior with infinite variance, IG$(2, b_i)$. In Models 2 and 4, where $\mathbf{AA}^T = \text{Diag}(\sigma_i^2)_{i=1}^5$, we again assign an IG$(2, b_i)$ to each of these variance parameters. The mean of this inverse-Gamma is determined by the b_i. To specify the hyperparameter b_i for each τ_i^2 and σ_i^2, we used nugget and partial sill estimates, respectively, from response-specific empirical semivariograms. In Models 3 and 4, $\mathbf{AA}^T$ is a full 5×5 covariance matrix, and therefore receives an inverse-Wishart prior with hyperparameters of five degrees of freedom and diagonal covariance matrix with diagonal elements drawn from the empirical semivariograms partial sill estimate.

The spatial decay, ϕ, and smoothness parameter, ν, used in the Matérn correlation function each receive a Uniform prior distribution. The combination of these parameters define the range of spatial dependence within the domain. If for instance $\nu = 0.5$, then (3.2) reduces to the familiar Exponential correlation function $\rho(\mathbf{s}, \mathbf{s}'; \phi, \nu) = \exp(-\phi\|\mathbf{s} - \mathbf{s}'\|)$ and the effective range (i.e., the distance at which the correlation drops to 0.05) is determined by $-\log(0.05)/\phi$. By allowing ν and ϕ to vary, the Matérn correlation function will produce a large interval of possible effective range values. Although we are interested in providing vague prior distributions, we want to set the support of ν and ϕ such that they allow for a reasonable effective range estimate. The maximum distance between any two plots in the BEF is 4,704.38 meters; therefore, we choose $\nu \sim U(0.1, 1.5)$ and $\phi \sim U(0.001, 0.1)$ which allows for an effective spatial range between about 10 and 4,750 meters. Obviously, other support on priors for ν and ϕ will produce a comparable interval for effective range; however, our previous experience suggests that these are reasonable priors.

As stated in Section 5.1, we assume that the candidate models share a common set of regressors. These regressors were chosen by backward elimination performed independently for each of the five response variables.

Programmatically, posterior sampling followed Section 4.1. For each sample, we used a single Metropolis–Hastings block-update of components in $\mathscr{A}$, $\boldsymbol{\theta}$, and Ψ with a multivariate normal proposal density. A Gibbs step then followed to update $\boldsymbol{\beta}$. Because all parameters in $\boldsymbol{\theta}$ and $\Psi > 0$ we actually update $\log(\phi)$, $\log(\nu)$, and $\log(\mathrm{Diag}(\tau_i^2)_{i=1}^5)$, then exponentiated each for use in the target likelihood. Similarly, we update $\mathbf{A}$ then calculate $\mathbf{K}$. Therefore, each parameter's Jacobian was required in the target likelihood.

We fit the five competing models to the data described in Section 2. The models were written in C++ and, being heavily dependent upon efficient matrix computations, leveraged the Intel® Math Kernel Library BLAS and LAPACK routines. For each of these models, three parallel MCMC chains were run for 20,000 iterations. The CODA package in R (*www.r-project.org*) was used to diagnose convergence by monitoring mixing, Gelman–Rubin diagnostics, autocorrelations, and cross-correlations. For each of the models, 5,000 iterations revealed sufficient mixing of the chains, so the remaining 45,000 samples (15,000 × 3) were retained for posterior analysis.

DIC was used to select the best candidate model to produce response-specific data layers for random spatial effects $E[\mathbf{W}|\ \mathrm{Data}]$, predicted random spatial effects $E[\mathbf{W}^*\ |\ \mathrm{Data}]$, and predicted biomass per hectare $E[\mathbf{Y}^*|\ \mathrm{Data}]$ with associated lower and upper 95% posterior predictive intervals. All interpolated surfaces presented here, using either the model plots or prediction plots, were produced using multilevel B-splines (Lee et al. 1997) computed with the `MBA` `R` package available at *www.r-project.org*.

Table 1. Model comparisons using the DIC criterion.

Model	Parameters	pD	DIC
Model 1	τ_m^2	35.2	8559.08
Model 2	$\nu, \phi, \sigma_m^2, \tau_m^2$	34.79	8542.76
Model 3	$\nu, \phi, \mathbf{A}, \tau_m^2$	34.63	8520.51
Model 4	$\nu_m, \phi_m, \sigma_m^2, \tau_m^2$	33.84	8535.98
Model 5	$\nu_m, \phi_m, \mathbf{A}, \tau_m^2$	34.69	8505.20

5.4 RESULTS

Table 1 provides pD and DIC scores for the competing models. Foremost, this table shows that the addition of spatial effects decreases DIC. The nearly constant estimates of effective number of parameters, pD, suggests that increased complexity results in shrinkage among the regressors and/or spatial process parameters. Decreased DIC scores in Models 3 and 5, over Models 2 and 4, support modeling of the conditional covariance among the response variables (i.e., conditional on the regressors). Holding the spatial variance constant, the nonseparable Models 4 and 5 perform better than the separable models. This suggests that the response variables exhibit different trends in conditional spatial dependence. Based solely on DIC, Model 5 provides the best model fit and therefore serves in subsequent analysis.

Estimates for the posterior distribution of each regressor's parameter are presented in Table 2. The sign and magnitude of estimates are consistent with univariate estimates found in the initial step-wise selection procedure. The credible intervals identify several regressors that contribute significantly to explaining variation in species-specific biomass per hectare. Because our focus is on optimal model selection, and not on understanding the functional relationship between the spectral variables and the response, we do not attempt to interpret these coefficients. We will point out that the signs on the elevation and slope coefficients are consistent with site conditions generally associated with the occurrence of these species and also agree with trends depicted in Figure 2. For instance, Eastern hemlock is typically found in lower elevations on moist soils, which corresponds to a negative SLOPE coefficient and high Eastern hemlock biomass values on shallow slopes, referring to Figures 1 and 2.

We now turn to estimates of the spatial process and measurement error parameters. The first block of parameters in Table 3 provides estimates for the elements in the square root of the 5 × 5 cross-covariance matrix. The subscripts on these parameters identify the species' variance or covariance. It is more instructive to convert the $\mathbf{AA}^T$ covariance matrix to a correlation matrix (5.1). This matrix provides a summary of the posterior conditional correlation among response variables, where rows and columns correspond to the species

Table 2. Percentiles of the posterior distribution of each regressor in Model 5. Each block of regressors correspond to one of the five response variables.

Parameters	50% (2.5%, 97.5%)	Parameters	50% (2.5%, 97.5%)
	BE Model		RM Model
Intercept	−480.75 (−747.52, −213.54)	Intercept	158.94 (16.92, 295.91)
ELEV	0.17 (0.09, 0.24)	ELEV	−0.07 (−0.14, 0.00)
AprTC2	1.72 (0.69, 2.74)	SLOPE	−1.76 (−2.75, −0.77)
AprTC3	−1.00 (−1.93, −0.06)	AprTC2	−0.87 (−1.42, −0.30)
AugTC1	3.39 (2.25, 4.51)	AugTC3	1.30 (0.43, 2.14)
AugTC3	1.45 (−0.25, 3.19)	OctTC2	−0.95 (−1.61, −0.29)
OctTC2	−0.77 (−1.83, 0.25)		SM Model
	EH Model	Intercept	−97.71 (−191.86, 0.62)
Intercept	−170.85 (−364.6, 21.45)	SLOPE	1.11 (0.48, 1.74)
SLOPE	−0.95 (−1.69, −0.17)	AugTC2	1.05 (0.71, 1.37)
AprTC1	2.08 (0.54, 3.66)	AugTC3	−0.44 (−1.1, 0.21)
AprTC2	−0.87 (−1.85, 0.13)		YB Model
AprTC3	1.75 (0.38, 3.13)	Intercept	−174.62 (−308.22, −29.63)
AugTC2	−0.65 (−1.11, −0.18)	ELEV	0.08 (0.01, 0.13)
AugTC3	1.54 (0.60, 2.51)	SLOPE	0.01 (−0.76, 0.82)
OctTC1	−1.74 (−2.76, −0.73)	AugTC1	0.27 (−0.26, 0.77)
OctTC2	1.55 (0.63, 2.45)	AugTC3	1.37 (0.47, 2.21)
OctTC3	−1.27 (−2.24, −0.31)		

BE, EH, RM, SM, YB and matrix elements are the 50% (2.5%, 97.5%) percentiles.

$$\left|\begin{array}{ccccc} 1 & & & & \\ 0.16(0.13, 0.21) & 1 & & & \\ -0.20(-0.23, -0.15) & 0.45(0.26, 0.66) & 1 & & \\ -0.20(-0.22, -0.17) & -0.12(-0.16, -0.09) & -0.48(-0.52, -0.41) & 1 & \\ 0.07(0.04, 0.08) & 0.22(0.20, 0.25) & 0.03(0.00, 0.09) & 0.01(-0.03, 0.03) & 1 \end{array}\right| \tag{5.1}$$

It is important to keep in mind that the cross-covariance, or cross-correlation, matrix captures association among the response variables conditional on the regressors. Therefore, it is best to interpret this correlation matrix in conjunction with the surface of random spatial effects from (4.3), depicted in Figure 3. Several significant correlations in (5.1), and corresponding spatial trends in Figure 3, support the earlier results suggesting that the regressors alone do not adequately account for extraneous variation in biomass per hectare. The significant correlations in (5.1) indicate strong spatial dependence among the five response variables, and corroborates the better performance of Model 5 in terms of DIC scores as seen in Table 1.

Returning to Table 3, the next block of parameters captures the measurement or pure error, $\Psi = \mathrm{Diag}(\tau_i^2)_{i=1}^5$. These values suggest that for all species there is a relatively large portion of variation not explained by the regressors or the spatial process. This seems especially true for red maple (RM) and sugar maple (SM). A more exhaustive set of covariates would help to reduce this variation further. However, we find the spatial variance components for American beech (BE), eastern hemlock (EH), and yellow birch (YB), that is, the

Table 3. Percentiles of the posterior distribution of variance and spatial range parameters from Model 5.

Parameters	Estimates: 50% (2.5%, 97.5%)
$\mathbf{K}_{\mathrm{BE}}$	1,969.01 (1,719.37, 2,446.00)
$\mathbf{K}_{\mathrm{BE,EH}}$	122.21 (112.07, 145.78)
$\mathbf{K}_{\mathrm{BE,RM}}$	−132.35 (−160.12, −106.93)
$\mathbf{K}_{\mathrm{BE,SM}}$	−98.21 (−117.39, −78.84)
$\mathbf{K}_{\mathrm{BE,YB}}$	43.73 (28.10, 71.33)
$\mathbf{K}_{\mathrm{EH}}$	312.64 (199.31, 495.09)
$\mathbf{K}_{\mathrm{EH,RM}}$	113.33 (82.99, 159.39)
$\mathbf{K}_{\mathrm{EH,SM}}$	−24.70 (−32.23, −16.69)
$\mathbf{K}_{\mathrm{EH,YB}}$	61.02 (55.83, 68.77)
$\mathbf{K}_{\mathrm{RM}}$	247.96 (118.83, 490.38)
$\mathbf{K}_{\mathrm{RM,SM}}$	−84.98 (−129.39, −49.63)
$\mathbf{K}_{\mathrm{RM,YB}}$	6.63 (−0.76, 12.08)
$\mathbf{K}_{\mathrm{SM}}$	124.36 (85.97, 158.89)
$\mathbf{K}_{\mathrm{SM,YB}}$	0.76 (−5.69, 4.73)
$\mathbf{K}_{\mathrm{YB}}$	265.31 (149.58, 359.64)
τ^2_{BE}	180.43 (142.24, 250.46)
τ^2_{EH}	423.18 (340.81, 491.75)
τ^2_{RM}	762.27 (653.28, 956.85)
τ^2_{SM}	728.21 (518.88, 938.64)
τ^2_{YB}	460.77 (385.36, 503.38)
ϕ_1	0.01070 (0.00896, 0.01498)
ϕ_2	0.00407 (0.00305, 0.00511)
ϕ_3	0.00441 (0.00332, 0.00646)
ϕ_4	0.00889 (0.00590, 0.01152)
ϕ_5	0.00981 (0.00346, 0.01238)
ν_1	0.45 (0.34, 0.67)
ν_2	0.46 (0.26, 0.61)
ν_3	0.67 (0.30, 0.74)
ν_4	0.31 (0.26, 0.42)
ν_5	0.55 (0.42, 0.68)

diagonal elements in $\mathbf{K}(\mathbf{0};\boldsymbol{\theta})$, do explain a substantial portion of the total variance. To be precise, we computed the posterior medians of the ratio $[\mathbf{K}(\mathbf{0};\boldsymbol{\theta})]_{ii}/([\mathbf{K}(\mathbf{0};\boldsymbol{\theta})]_{ii} + \tau_i^2)$ for $i = 1, \ldots, 5$ (corresponding to the five response variables) to be 0.92, 0.42, 0.25, 0.15, and 0.37, respectively.

The last two blocks in Table 3 provide point and credible interval estimates for the spatial decay and smoothness parameters present in the spatial correlation function parameters, ϕ_k and ν_k. Based on these parameters' credible interval we conclude that the Uniform

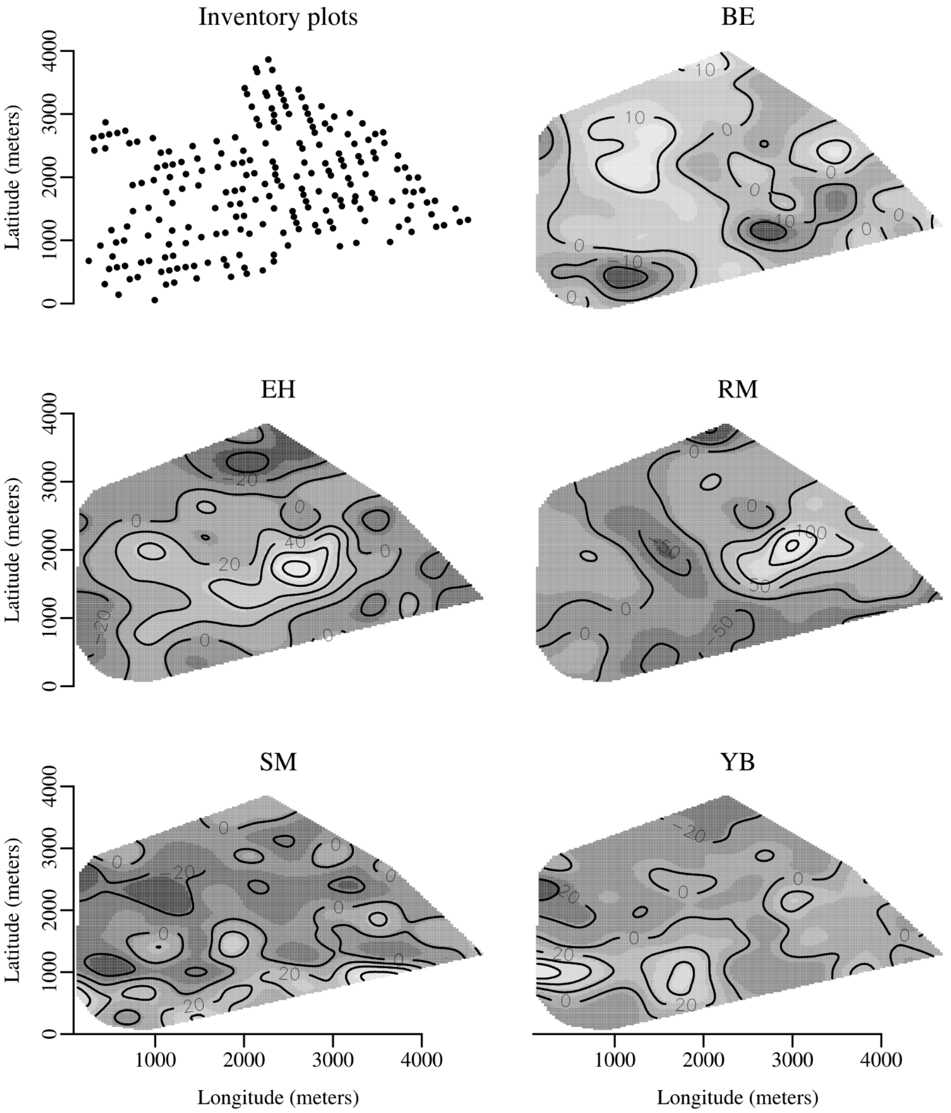

Figure 3. Interpolation of the recovered random spatial effects of biomass per hectare, $E[\mathbf{W}|\text{ Data}]$, by species from Model 5.

priors' support was sufficiently vague and did not influence the course of the chains (i.e., the credible intervals are well within the defined support). It is best to use these estimates to solve the Matérn correlation function for the effective range of spatial dependence (i.e., the distance at which $\rho = 0.05$). Table 4 provides the median and 95% credible interval for each process effective range. Considering that the inventory plots are laid out on a 200×100 meter grid and the estimated effective range bounds, we conclude that observations made on a given inventory plot are not independent from those of neighboring plots. This spatial dependence must be considered in prediction. Again referring to Table 4, there is a

Table 4. Distance in meters at which the spatial correlation drops to 0.05 for each of the underlying spatial processes. Distance calculated by solving the Matérn correlation function for d using $\rho = 0.05$ and process specific ϕ and ν parameters estimates from Model 5.

Process	Estimates: 50% (2.5%, 97.5%)
$\tilde{W}_1(\mathbf{s})$	270.84 (200.15, 334.52)
$\tilde{W}_2(\mathbf{s})$	697.47 (466.23, 998.97)
$\tilde{W}_3(\mathbf{s})$	756.50 (504.08, 954.28)
$\tilde{W}_4(\mathbf{s})$	275.10 (207.13, 395.49)
$\tilde{W}_5(\mathbf{s})$	314.03 (253.71, 856.79)

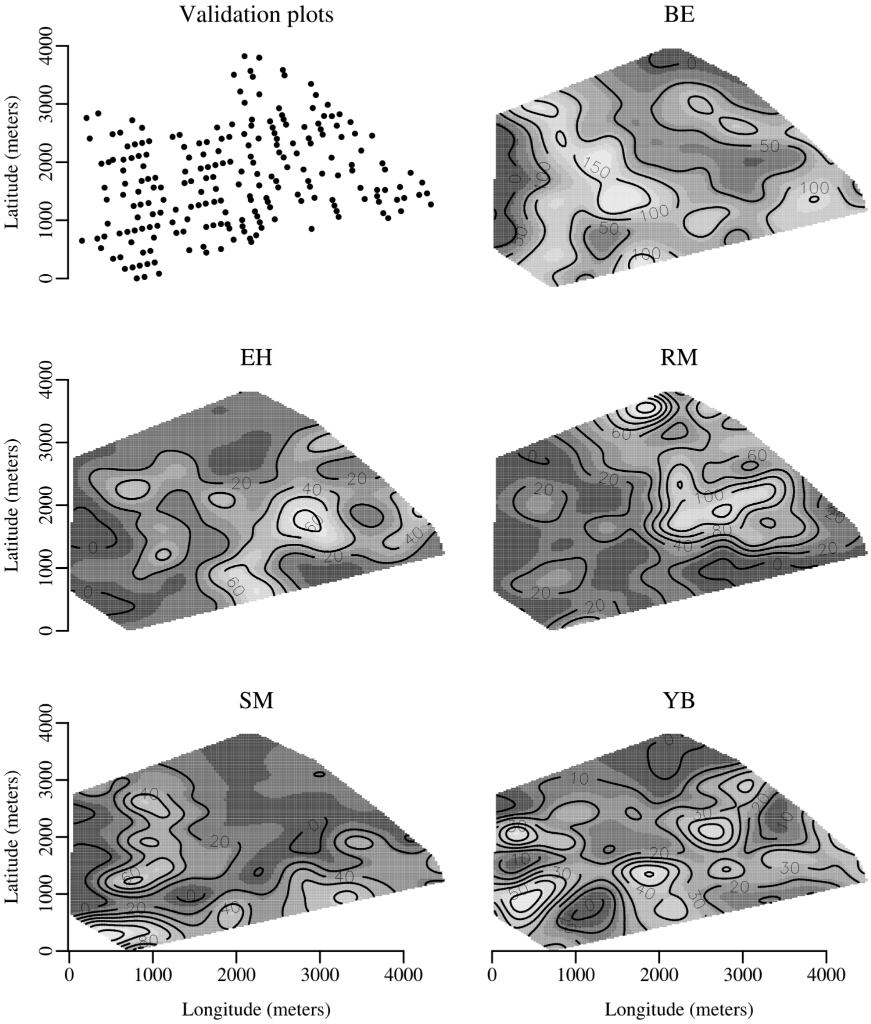

Figure 4. Interpolation of metric tons of biomass per hectare by species measured on inventory plots across the BEF. This set of 218 inventory plots was used for model validation.

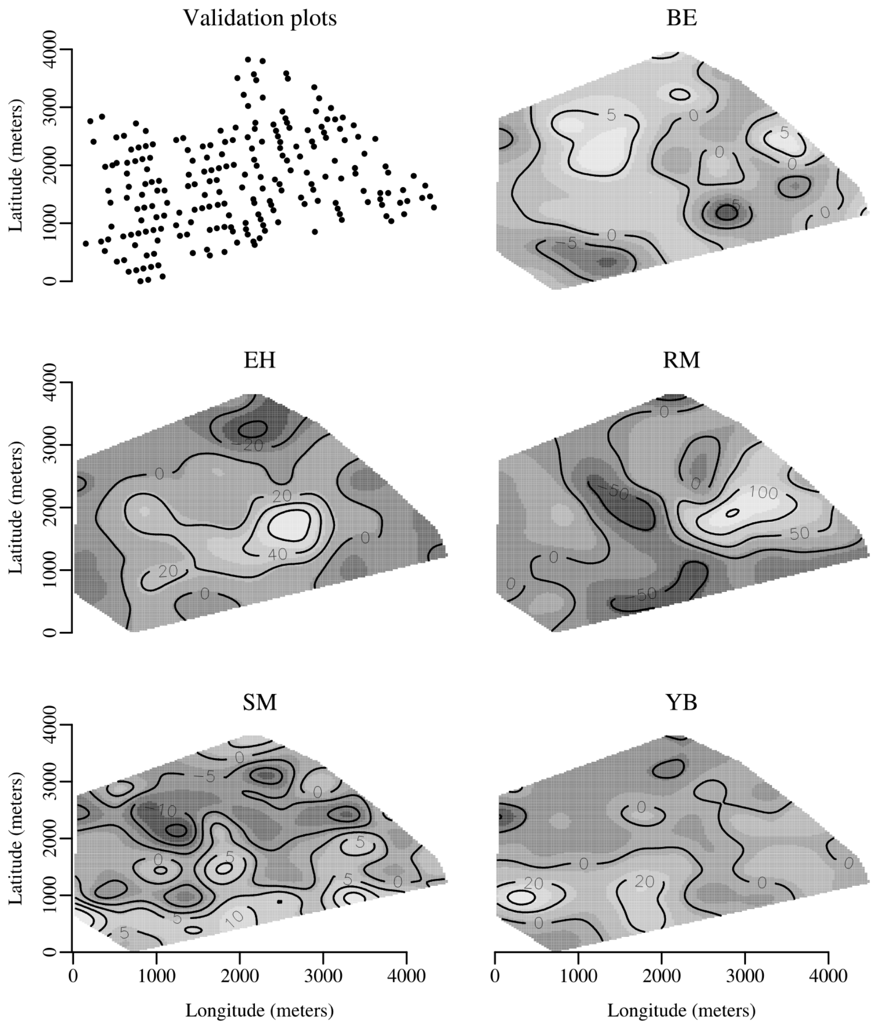

Figure 5. Interpolation of the predicted random spatial effects of biomass per hectare, $E[\mathbf{W}^* \mid \text{Data}]$, by species from Model 5.

relatively large disparity in the range of the underlying $\tilde{\mathbf{W}}(\mathbf{s})$ driving the spatial dependence among several species. Specifically, $\tilde{W}_1(\mathbf{s})$ and $\tilde{W}_4(\mathbf{s})$ have small effective spatial range compared with $\tilde{W}_2(\mathbf{s})$ or $\tilde{W}_3(\mathbf{s})$. This result strongly supports the use of the nonseparable models that allow for different rates of spatial decay.

Our central purpose in fitting this model was to gain access to the multivariate posterior predictive distribution of biomass per hectare of any set of newly observed points across the BEF (e.g., the set of validation points Figure 4). Given the set of validation points $\{\mathbf{s}_{0i}\}_{i=1}^{n^*}$, where $n^* = 218$ and the posterior samples $\{\Omega^{(l)}\}_{l=1}^{L} \sim P(\Omega| \text{ Data})$, where $L = 45{,}000$, we might first use (4.4) to compute and map (Figure 5) the posterior predictive distribution $P(\mathbf{W}^* \mid \text{Data})$ (recall $\mathbf{W}^* = \mathscr{A}\tilde{\mathbf{W}}^*$). Because of the strong conditional spatial dependence exhibited by the response variables, this surface of predicted spatial random effects resembles the recovered spatial random effects surface (Figure 3).

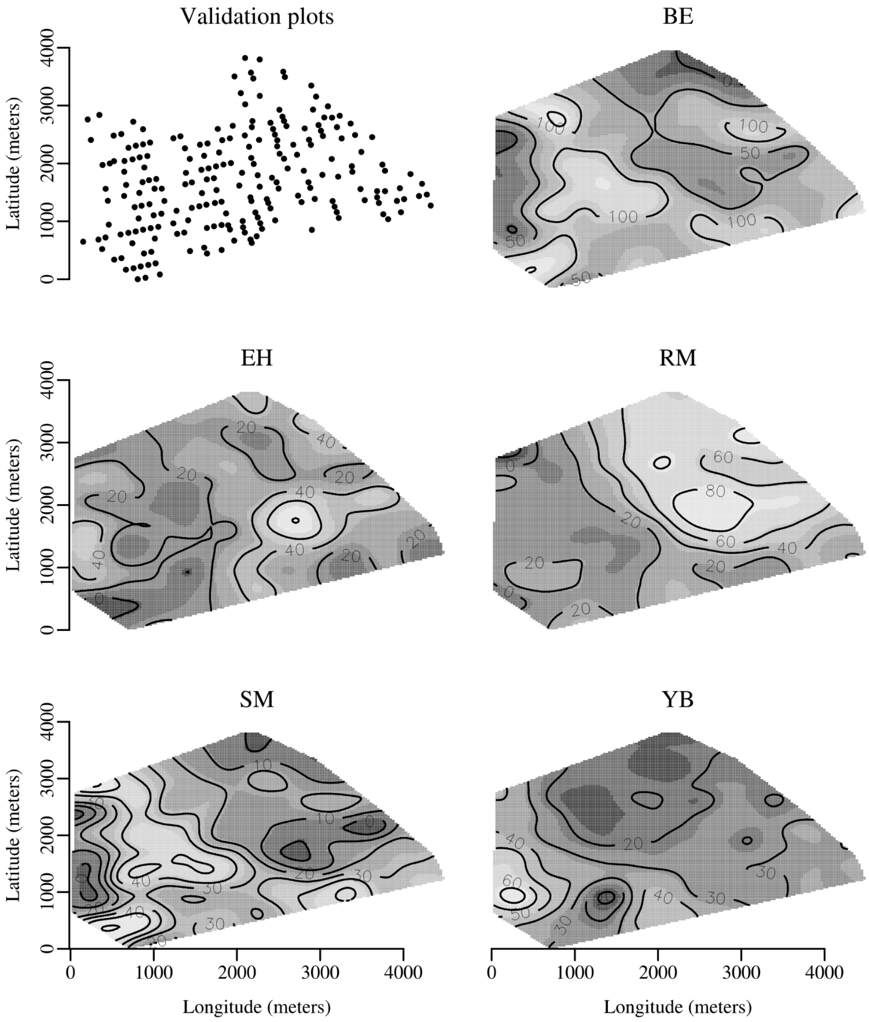

Figure 6. Interpolation of the mean of the posterior predictive sample of biomass per hectare, $E[\mathbf{Y}^* \mid \text{Data}]$, by species from Model 5.

Prediction of the validation points, $P(\mathbf{Y}^* \mid \text{Data})$, follows (4.5). Figure 6 provides the mean of each species' metric tons per hectare predictive distribution. Except for a few departures and inherent smoothing, the predicted surfaces generally follow the observed surfaces (Figure 4). Referring to Figures 4 and 6, we again see American beech (BE) following the observed trend of greater biomass volume in higher elevations and on moderate slopes. Eastern hemlock (EH) dominates shallow slopes, but is overestimated on the higher elevations which suggests that ELEV should have been forced to stay in the model. Red maple (RM) follows the observed surface, except for where there is a paucity of validation points in the upper elevations. Sugar maple (SM) biomass volume seems to be overestimated at lower elevations and in the western portion of the BEF. Finally, yellow birch (YB) generally follows the observed trends but is oversmoothed.

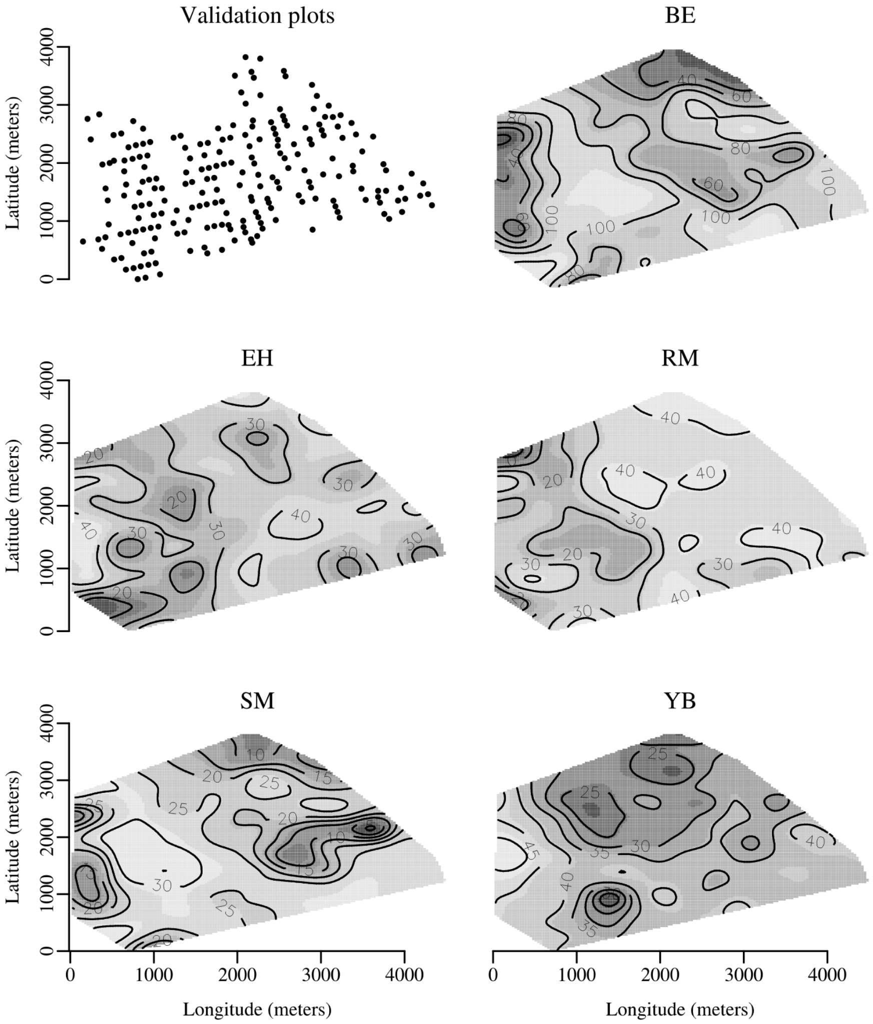

Figure 7. Interpolation of the range between the lower and upper 95% posterior predictive intervals of biomass per hectare, by species from Model 5.

Access to each species' full posterior predictive distribution affords great flexibility in summarizing the uncertainty in the spatially explicit predictions. This is a key strength of the sampling-based framework. Based on these distributions, Figure 7 depicts the range between the 0.025 and 0.975 quantiles (i.e., $q_U - q_L$ of $P(q_L < Y^* < q_U|\, \text{Data}) = 1 - \alpha$). However, any percentile or function of the predictive distribution can be mapped. Beyond describing the uncertainty in estimates of biomass per hectare, these error surfaces can reveal missing regressors and regions within the domain with insufficient observations.

6. DISCUSSION

This article focused on the development of a general Gaussian template for multivariate spatial regression models with Gaussian responses that arise in the biological and environmental sciences. Through our proposed template, we investigated the multivariate BEF forest biomass dataset using models with increasing richness in correlation structures. Our example demonstrated that simple covariance structures, such as those that ignore covariances between the response variables or impose common spatial ranges for each of the variables, yield poorer fits to the data. The parameter estimates from the "best" model were used to construct a predictive surface of forest biomass for each major species on the BEF. Most importantly, access to the full multivariate posterior predictive distribution permitted mapping of uncertainty in these predictions. Data layers such as these can serve as input variables to subsequent models used to help us better understand forest carbon dynamics in the northeastern United States.

We foresee future work in several directions. On the practical front, we intend to conduct more in-depth investigations into the statistical versus practical criterion for model selection, such as comparing DIC with Root Mean Square Error (RMSE) calculated using the validation set. In the BEF dataset, although DIC selects Model 5 we observe rather marginal improvements in RMSE calculated within the validation set—from a score of about 58 for Model 2 (the worst spatial model) to 50 for Model 5 (the best spatial model). We expect this reduction to be enhanced by higher cross-correlations between the variables (the highest cross-correlation for our current example was approximately 0.45) that would allow learning through the off-diagonal elements.

On the methodological side, we envision richer structures for the Ψ matrix, which we specified as diagonal here. These present data did not yield good convergence with a general inverted Wishart prior on Ψ. This is because with one multivariate observation from each location the data is unable to identify such rich structures in Ψ. We could, however, consider replicated multivariate measurements from each location that would assist in accommodating such structures which could represent nonspatial residual correlation. In such replicated settings we could also estimate nonstationary multivariate processes with the cross-covariance varying across space and being captured by the space-varying linear transformation $\mathbf{A}(\mathbf{s})$. As a further step one could devise strategies of modeling $\mathbf{A}(\mathbf{s})\mathbf{A}^T(\mathbf{s})$ using a matrix-variate spatial process.

Also, in certain contexts one may consider relaxing the assumption about the nonsingularity of $\mathbf{A}$. This may be relevant when a very large dimensional spatial processes needs to be projected onto a span of a smaller number of independent processes. For example, the m-variate process $\mathbf{W}(\mathbf{s})$ may be related to a p-variate process $\tilde{\mathbf{W}}(\mathbf{s})$ $(p < m)$ as $\mathbf{W}(\mathbf{s}) = \mathbf{A}\widetilde{\mathbf{W}(\mathbf{s})}$, where $\mathbf{A}$ is now an $m \times p$ matrix. In fact, often in practice two or three components for $\tilde{\mathbf{W}}(\mathbf{s})$ are able to capture the underlying spatial variation. However, now $\mathbf{A}\mathbf{A}^T$ is rank-deficient and the inverse Wishart prior is precluded. One may parameterize this $\mathbf{A}\mathbf{A}^T$ in terms of its Givens angles and eigenvectors and assign priors to them. Daniels and Kass (1999) provided a detailed theoretical investigation into such priors.

For the models that we explored, the off-diagonal elements of $\mathbf{K}$ and Ψ only contribute to learning if the variance between response variables is a linear function. What if there is some nonlinear association among response variables' variance? We might try linearizing transformations; however, a more fruitful approach would allow general functional relationships within the cross-covariance and cross-measurement error matrices. Then our task of guaranteeing a positive definite Σ becomes more difficult; however, the reward in improved precision could be substantial. Browne (2006) discussed how we might ensure the correct condition of the final variance matrix.

Finally, to make our methodology more far-reaching in its usage, we need to recognize that the computational burden for implementing our template will explode with a large number of locations. This is known as the so-called "big-N" problem in spatial statistics and is an area of active research. Strategies for addressing this problem involve representing the spatial process $\mathbf{W}(\mathbf{s})$ over a smaller set of representative locations (called knots) (see, e.g., Banerjee et al. 2004). These methods should also be applicable to spatiotemporal settings where the computational burden is considerably increased with the added dimension. It is also worth noting that Zhang (2006) considered an EM algorithm for maximum likelihood estimation (essentially treating $\mathbf{W}(\mathbf{s})$ as "missing" spatial effects) for a spatially static linear model of coregionalization. The EM algorithm has some desirable properties though it does not incorporate model uncertainty, hence we did not pursue this approach here. However, Zhang (2006) derived explicit closed-form expressions and shows that the resulting matrix estimates of $\mathbf{A}$ are positive semidefinite. Therefore it is potentially useful for dealing with a high-dimensional multivariate process given a large number of spatial locations.

ACKNOWLEDGMENTS

The work of the first author was supported by the NASA Headquarters Earth System Science Fellowship Grant NGT5, USDA Forest Service Forest Inventory and Analysis Program, and National Science Foundation grant DMS-0706870. The work of the second author was supported by the National Institutes of Health grant 1R01CA112444-01 and National Science Foundation grant DMS-0706870. The authors are grateful to three anonymous reviewers for their many helpful comments and suggestions.

[Received August 2006. Revised July 2007.]

REFERENCES

Banerjee, S., Carlin, B.P., and Gelfand, A.E. (2004), *Hierarchical Modeling and Analysis for Spatial Data*, Boca Raton, FL: Chapman and Hall/CRC Press.

Banerjee, S., and Finley, A.O. (2007), "Bayesian Multiresolution Modeling of Spatially Replicated Data," *Journal of Statistical Planning and Inference*, 137, 3193–3205.

Banerjee, S., and Johnson, G.A. (2006), "Coregionalized Single- and Multi-resolution Spatially-Varying Growth Curve Modelling with Applications to Weed Growth," *Biometrics*, 62, 864–876.

Browne, W.J. (2006), "MCMC Algorithms for Constrained Variance Matrices," *Computational Statistics and Data Analysis*, 50, 1655–1677.

Carlin, B.P., and Louis, T.A. (2000), *Bayes and Empirical Bayes Methods for Data Analysis*, Boca Raton, FL: Chapman and Hall

Chilés, J.P., and Delfiner, P. (1999), *Geostatistics: Modelling Spatial Uncertainty*, New York: Wiley.

Cressie, N.A.C. (1993), *Statistics for Spatial Data* (2nd ed.), New York: Wiley.

Daniels, M.J., and Kass, R.E. (1999), "Nonconjugate Bayesian Estimation of Covariance Matrices and its use in Hierarchical Models," *Journal of the American Statistical Association*, 94, 1254–1263.

Gelfand A.E., Schmidt, A., Banerjee S., and Sirmans, C.F. (2004), "Nonstationary Multivariate Process Modelling through Spatially Varying Coregionalization," *Test*, 13, 263–312.

Gelman, A., Carlin, J.B., Stern, H.S., and Rubin, D.B. (2003), *Bayesian Data Analysis* (2nd ed.), Boca Raton, FL: Chapman and Hall/CRC Press.

Harville, D.A. (1997), *Matrix Algebra from a Statistician's Perspective*, New York: Springer.

Homer, C., Huang, C., Yang, L., Wylie, B., and Coan, M. (2004), "Development of a 2001 National Land-cover Database for the United States," *Photogrammetric Engineering and Remote Sensing*, 70, 829–840.

Huang, C., Wylie, B., Homer, C., Yang, L., and Zylstra, G. (2002), "Derivation of a Tasseled Cap Transformation Based on Landsat 7 At-Satellite Reflectance," *International Journal of Remote Sensing*, 8, 1741–1748.

Hudak, A.T., Lefsky, M.A., and Cohen, W. B. (2002), "Integration of Lidar and Landsat ETM+ Data for Estimating and Mapping Forest Canopy Height," *Remote Sensing of Environment*, 82, 397–416.

Katila, M., and Tomppo, E. (2001), "Selecting Estimation Parameters for the Finnish Multi-source National Forest Inventory," *Remote Sensing of Environment*, 76, 16–32.

Lappi, J. (2001), "Forest Inventory of Small Areas Combining the Calibration Estimator and a Spatial Model," *Canadian Journal of Forest Research*, 31, 1551–1560.

Lee, S., Wolberg, G., and Shin, S.Y. (1997), "Scattered Data Interpolation with Multilevel B-splines," *IEEE Transactions on Visualization and Computer Graphics*, 3, 229–244.

Majumdar, A., and Gelfand, A.E. (2006), "Spatial Modeling for Multivariate Environmental Data Using Convolved Covariance Functions," Technical Report Institute of Statistics and Decision Sciences, Duke University, Durham, NC.

McRoberts, R.E., Nelson, M.D., and Wendt, D.G. (2002), "Stratified Estimation of Forest Area using Satellite Imagery, Inventory Data, and the *k*-Nearest Neighbors Technique," *Remote Sensing of Environment*, 82, 457–468.

Spiegelhalter, D.J., Best, N.G., Carlin, B.P., and van der Linde, A. (2002), "Bayesian Measures of Model Complexity and Fit" (with discussion), *Journal of the Royal Statistical Society*, Series B, 64, 583–639.

Stein, M.L. (1999), *Interpolation of Spatial Data: Some Theory for Kriging*, New York: Springer.

Tomppo, E. (1991), "Satellite Imagery-Based National Forest Inventory of Finland," *International Archives of Photogrammetry and Remote Sensing*, 28, 419–424.

Tomppo, E., and Halme, M. (2003), "Using Coarse Scale Forest Variables as Ancillary Information and Weighting of Variables in k-NN Estimation: A Genetic Algorithm Approach," *Remote Sensing of Environment*, 92, 1–20.

Trotter, C.M., Dymond, J.R., and Goulding, C.J. (1997), "Estimation of Timber Volume in a Coniferous Plantation Forest using Landsat TM," *International Journal of Remote Sensing*, 18, 2209–2223.

Wackernagel, H. (2003), *Multivariate Geostatistics: An Introduction with Applications* (3rd ed.), Berlin: Springer-Verlag.

Zhang, H. (2006), "Maximum-Likelihood Estimation for Multivariate Spatial Linear Coregionalization Models," *Environmetrics*, 18, 125–139.